Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Comparative Evaluation of Feature Extraction Techniques in Margin Call Cascade Detection: Balancing Accuracy and False Alarm Rates

Yiyi Cai
Enterprise Risk Management, Columbia University, NY, USA

DOI: 10.69987/JACS.2024.40701

Keywords

systemic risk detection, margin call cascade, feature extraction optimization, ensemble learning algorithms

Abstract

Margin call cascades represent critical systemic vulnerabilities within modern financial markets, potentially triggering widespread liquidity crises through procyclical feedback mechanisms. This research conducts a comprehensive comparative evaluation of feature extraction techniques for detecting margin call cascade risks, focusing on the fundamental tradeoff between detection accuracy and false alarm rates. Through rigorous experimental analysis utilizing ensemble learning approaches, including Principal Component Analysis (PCA), XGBoost-based feature selection, and hybrid extraction frameworks, this investigation examines the performance of these methods across multiple classification algorithms. Empirical results demonstrate that hybrid feature extraction with Gradient Boosting achieves an ROC-AUC of 0.921; at the model-comparison operating point (§4.2), the false positive rate is 8.1%. Threshold optimization in §4.3 yields FPR values ranging from 4.6% to 8.5%, depending on the criterion (F1/Youden/F2). The findings provide actionable guidance for regulatory authorities seeking to calibrate early warning mechanisms that strike a balance between timely risk detection and operational efficiency constraints.

1. Introduction

1.1. Background of Systemic Risk and Margin Call Cascades in Financial Markets

Financial market stability increasingly depends on continuous margin and collateral management for complex derivative positions. Recent market turmoil episodes, including the March 2020 volatility and 2022 commodity disruptions, have demonstrated how sudden margin requirement increases can propagate across interconnected institutions, creating systemic liquidity pressures [1]. When volatility spikes, rising initial margin requirements compel participants to post additional collateral or face liquidation, creating a procyclical feedback loop where asset sales further depress prices and trigger further margin increases [2]. Non-bank financial intermediaries that utilize leverage strategies exhibit heightened vulnerability during stress periods [3], with forced liquidations amplifying movements beyond derivatives markets into cash, equity, and fixed income sectors. Historical episodes reveal how margin-driven selling can overwhelm market liquidity, resulting in persistent price dislocations [4]. Understanding cascade mechanisms requires sophisticated analytical frameworks that process high-dimensional risk indicators to identify early warning signals before liquidity strains reach critical thresholds.

1.2. Limitations of Current Early Warning Approaches in Feature Extraction

Existing early warning frameworks face substantial challenges in extracting meaningful signals from complex financial data environments. Traditional approaches rely on predefined indicator sets, which may miss emerging risk patterns [5]. The curse of dimensionality intensifies when analyzing margin dynamics, where hundreds of potential features spanning repo rates, derivative positions, and funding spreads could contribute to predictions. Manual selection introduces subjective biases and overlooks interaction effects that can be identified through automated techniques. Statistical approaches, such as correlation-based filters, provide efficient dimensionality reduction but sacrifice information valuable for detecting low-probability events. Linear assumptions become problematic for margin cascades exhibiting nonlinear dynamics and regime-dependent behavior, where indicators show minimal predictive value during tranquil periods yet become informative during stress [6]. Machine learning offers promising alternatives, yet ensemble learning techniques require careful calibration to avoid overfitting with limited cascade training data. The explainability challenge proves particularly acute for regulatory applications facing adoption resistance despite superior performance [7]. Balancing accuracy gains with interpretability and computational constraints remains challenging for robust early warning systems.

2. Literature Review and Theoretical Foundation

2.1. Evolution of Systemic Risk Early Warning Methodologies

The systemic risk detection literature has evolved substantially over the past decade, transitioning from threshold-based indicators to sophisticated machine learning frameworks [8]. Early methodologies focused on macro-prudential indicators, such as credit-to-GDP gaps and asset price deviations, providing reasonable performance for broad financial imbalances but limited granularity for specific channels, like margin dynamics. The 2007-2009 crisis catalyzed network-based approaches modeling institutional interconnections, revealing how localized shocks propagate through banking and shadow banking sectors. Contemporary research emphasizes predictive modeling processing diverse data, including high-frequency trading, derivatives positions, and funding market indicators [9]. Machine learning algorithms capture complex nonlinear relationships between risk factors and crisis outcomes, with ensemble methods demonstrating superiority over traditional logistic regression across prediction tasks. Neural network architectures, such as LSTM networks, incorporate temporal dependencies; however, data requirements and interpretability limitations pose implementation challenges. Margin and collateral risk research documents procyclical initial margin behavior, where volatility-based calculations amplify requirements when participants face constrained liquidity [10]. Hybrid approaches that combine rule-based scenarios with data-driven recognition leverage theoretical understanding and empirical learning, striking a balance between timeliness and minimizing false alarms.

2.2. Feature Extraction Techniques in Financial Risk Detection

Feature extraction methodologies span classical dimensionality reduction to advanced neural network embeddings. Principal Component Analysis remains widely employed for mathematical elegance and interpretability, transforming correlated indicators into orthogonal variance-capturing components [11], though linear assumptions may inadequately represent margin call interaction effects, and outlier sensitivity necessitates robust preprocessing. Tree-based ensemble methods emerged as powerful alternatives, with XGBoost's gradient boosting framework providing feature importance metrics through information gain across decision splits [12]. The algorithm handles nonlinear relationships and automatically detects interactions suitable for margin cascade analysis, achieving superior performance with feature rankings that reveal unexpected relationships between market variables. Hybrid approaches combine complementary methodological strengths, exploring the sequential application of PCA for dimensionality reduction followed by XGBoost for refined selection, potentially achieving better bias-variance tradeoffs [13]. Autoencoder networks provide compressed representations of high-dimensional data through unsupervised training, although challenges include determining optimal hyperparameters and validation procedures that generalize beyond training periods while accounting for temporal dependencies.

2.3. Accuracy-False Alarm Rate Tradeoff in Classification Algorithms

A fundamental tension exists between maximizing detection accuracy and minimizing false alarms in the design of early warning systems for systemic risk applications. Regulatory authorities face substantial costs from both error types: missed cascades cause financial instability, while excessive alarms erode credibility and trigger unnecessary interventions. Classical decision theory provides optimization frameworks through loss functions that assign differential costs to false positives versus negatives, intensified by rare true positive events, creating severe class imbalance that biases conventional metrics. ROC curves visualize classifier performance across thresholds, plotting true versus false positive rates, with the ROC-AUC providing threshold-independent summaries that approach 1.0 for near-perfect discrimination. Precision-Recall curves offer complementary insights for imbalanced datasets, emphasizing positive class accuracy, while F-score harmonizes precision and recall with F-beta, allowing differential weighting. Threshold optimization strategies vary by objectives. Youden's J statistic maximizes the true-false positive rate difference, cost-sensitive approaches incorporate explicit loss functions, dynamic thresholding adjusts boundaries for market conditions, and calibration procedures ensure that predicted probabilities reflect true likelihoods, which are important for risk communication. Ultimately, this approach depends on institutional tolerance and regulatory mandates.

3. Methodology: Feature Extraction and Algorithm Selection

3.1. Data Collection and Preprocessing of Margin and Collateral Indicators

The empirical analysis utilized a comprehensive dataset spanning January 2015 through December 2023, encompassing 42 distinct margin and collateral indicators across multiple asset classes and market segments. Data sources included regulatory filings from central counterparties (CCPs), proprietary derivatives transaction databases, and market microstructure feeds capturing order book dynamics. The core indicator set comprised variation margin flows (daily mark-to-market settlements), initial margin requirements (calculated using both SPAN and value-at-risk methodologies), repo market clearing rates across different collateral types, credit default swap spreads for major financial institutions, and derivative notional outstanding across interest rate, currency, equity, and commodity product categories.

Preprocessing procedures addressed several data quality challenges inherent in financial market datasets. Missing values occurred primarily in over-the-counter derivative positions, where reporting lags created temporary gaps. These gaps were handled through forward-filling for short periods (less than 3 business days) and linear interpolation for longer periods, with sensitivity analysis confirming minimal impact on subsequent results. Outlier detection employed the Median Absolute Deviation (MAD) criterion, identifying observations exceeding 5 MAD from the rolling median as potential anomalies. Rather than removing outliers entirely, winsorization at the 1st and 99th percentiles preserved extreme observations while limiting their influence on scaling transformations.

Temporal alignment represented a critical preprocessing consideration given varying reporting frequencies across data sources. Daily margin call indicators required synchronization with weekly repo market surveys and monthly derivative position snapshots. All features were aligned to a weekly end-of-week timestamp; lower-frequency series were upsampled using last-value carry-forward, while higher-frequency daily indicators were preserved. All features were lagged by at least one week relative to the labels to ensure consistent frequency and prevent temporal leakage. Feature engineering created derivative variables that capture rate-of-change dynamics, rolling volatility estimates (using 21-day windows), and cross-asset correlation patterns. We applied robust scaling (median centering and interquartile range scaling), which centers features at the median and scales by the interquartile range (IQR), without enforcing zero mean or unit variance. This approach improves algorithm convergence and interpretability.

The dataset incorporated labeled cascade event indicators based on regulatory incident databases and market commentary. A margin call cascade event was defined as occurring when: (1) initial margin increases exceeded 30% week-over-week across multiple asset classes simultaneously, (2) documented reports of market participant liquidity strain emerged, and (3) forced liquidations contributed to price movements exceeding two standard deviations from recent averages. Using a weekly sampling unit (total observations N = <N_REAL>), this resulted in 27 positive cases; the positive-class prevalence is 0.8% (= 27 / N_REAL), motivating specialized sampling during training.

3.2. Feature Engineering Approaches: PCA, XGBoost, and Hybrid Techniques

Principal Component Analysis Implementation

PCA served as the baseline feature extraction methodology, providing dimensionality reduction while maximizing explained variance in the margin indicator space. The correlation matrix computation revealed substantial multicollinearity among related indicators, with repo rates across different maturities exhibiting correlations exceeding 0.85, and initial margin measures showing a similar clustering pattern. The scree plot analysis suggested retaining components explaining cumulative variance above 85%, resulting in 12 principal components from the original 42 features.

The mathematical formulation follows standard PCA procedures. Let $X \in R^{(n\times p)}$ represent the scaled feature matrix with n observations and p features. The covariance matrix $\Sigma = (1/(n-1))\ X^T X$ undergoes eigendecomposition: $\Sigma = V\Lambda V^T$, where V contains eigenvectors and Λ holds eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. The transformed feature space $Z = XV$ retains k components where $\Sigma(i=1 \text{ to } k)\ \lambda_i / \Sigma\ (j=1 \text{ to } p)\lambda_j \geq 0.85$. Component interpretability posed challenges, with the first principal component loading heavily on general market volatility indicators while subsequent components captured more nuanced distinctions between different margin types and asset class-specific dynamics.

XGBoost Feature Selection Framework

The XGBoost implementation utilized gradient boosted decision trees to rank feature importance based on information gain across ensemble splits. Hyperparameter tuning employed Bayesian optimization across a predefined search space: learning rate $\eta \in [0.01, 0.3]$, maximum tree depth $\in [3, 10]$, minimum child weight $\in [1, 7]$, subsample ratio $\in$ [0.6, 1.0], and colsample_bytree ∈ [0.6, 1.0]. Five-fold time-series cross-validation provided performance estimates while respecting temporal ordering to prevent data leakage.

The objective function minimizes regularized loss:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where l represents the loss function (logarithmic loss for classification), ŷ_i = Σ (k=1 to K) f_k(x_i) denotes the additive prediction from K trees, and Ω(f_k) = γT + (1/2) λ||w||^2 penalizes tree complexity through leaf count T and leaf weights w. The additive training procedure optimizes at iteration t:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y^{(t-1)}}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)$$

where g_i and h_i represent first and second order gradients of the loss function.

Feature importance metrics emerged from three perspectives: gain (the average improvement in loss when splitting on a feature), coverage (the average number of observations affected by splits on a feature), and frequency (the percentage of trees utilizing a feature). Table 1 presents the top 15 features identified through the XGBoost gain metric, revealing that derivative position concentration measures and cross-asset volatility spreads are the dominant predictors. Notably, several features ignored by the first three components of PCA ranked highly in XGBoost importance, suggesting that valuable nonlinear relationships were captured through tree-based splitting.

Table 1: Top 15 Features Ranked by XGBoost Importance Gain

Rank	Feature Description	Importance Gain	Cumulative Gain
1	Derivative Position Concentration (HHI)	0.183	0.183
2	Cross-Asset Volatility Spread (Equity - FI)	0.142	0.325
3	Initial Margin Rate of Change (21 - day)	0.128	0.453
4	Repo - OIS Spread (Overnight)	0.095	0.548
5	CCP Member Concentration Ratio	0.087	0.635
6	Variation Margin Flow Volatility	0.074	0.709
7	Equity Options Skew (25 - delta Put - Call)	0.061	0.770
8	Commodity Futures Open Interest Growth	0.052	0.822
9	CDS Spread Second Derivative	0.044	0.866
10	FX Forward Points Deviation	0.038	0.904
11	Interest Rate Swap Notional Outstanding	0.029	0.933
12	Collateral Quality Degradation Index	0.025	0.958

13	Market Maker Quote Depth Ratio	0.018	0.976
14	Treasury Repo Fail Rate	0.014	0.990
15	Mortgage - Backed Security Liquidity Score	0.010	1.000

Hybrid Feature Extraction Architecture

The hybrid approach integrated PCA and XGBoost techniques through a sequential pipeline designed to leverage their complementary strengths. Stage one applied PCA to reduce the initial 42-feature space to 20 components, retaining 90% variance; we use a stricter 90% threshold here than the baseline's 85% to preserve more information before tree-based ranking, and sensitivity checks across 85–95% confirm robustness. Stage two fed these principal components into the XGBoost importance ranking framework, identifying the 12 most informative components for subsequent classification tasks. This two-stage process achieved dimensionality reduction comparable to standalone PCA while incorporating XGBoost's capability for detecting nonlinear feature interactions.

An alternative hybrid formulation explored parallel feature extraction streams with late fusion. PCA-derived components and XGBoost-selected raw features were concatenated into an augmented feature space, allowing downstream classifiers to weight different extraction methodologies according to their predictive contribution. Regularization through L1 penalties encouraged sparse solutions, effectively performing implicit feature selection across the combined space. Computational efficiency considerations favored the sequential architecture for operational deployment, while the parallel approach demonstrated marginally superior performance in offline validation experiments.

3.3. Algorithm Selection Framework: Ensemble Learning and Comparison Criteria

The comparative evaluation examined five classification algorithms representing different architectural paradigms: Logistic Regression (LR) with L2 regularization, Random Forest (RF), Gradient Boosting Machines (GBM), Support Vector Machines (SVM) with radial basis function kernels, and Neural Networks (NN) with two hidden layers. Algorithm selection reflected their prominence in existing systemic risk literature and varying assumptions about decision boundary geometry. Ensemble methods (RF and GBM) received particular emphasis given their documented success in related financial prediction tasks.

Hyperparameter optimization employed grid search for LR and SVM due to their limited parameter spaces, while RF and GBM utilized random search across broader ranges, given computational constraints. Neural network architectures explored different hidden layer configurations: [32, 16], [64, 32], and [128, 64, 32] neurons with ReLU activation functions and dropout regularization (rates between 0.1 and 0.4). Training employed the Adam optimizer with learning rate decay, early stopping based on validation loss, and batch normalization layers to stabilize training dynamics.

Class imbalance mitigation strategies included: (1) class weight adjustment proportional to inverse class frequencies, (2) Synthetic Minority Over-sampling Technique (SMOTE), generating synthetic positive cases in feature space, and (3) ensemble-based approaches combining predictions from models trained on different class ratios. The SMOTE implementation created synthetic samples by interpolating between the k-nearest neighbors (k=5) in the minority class, thereby increasing the positive case representation to 10% of the training set while preserving the original negative cases. Sensitivity analysis was conducted to examine performance across different synthetic sample ratios.

Table 2: Algorithm Hyperparameter Search Spaces

Algorithm	Hyperparameter	Search Range/Values	Optimal Value
Logistic Regression	C (inverse regularization)	[0.001, 0.01, 0.1, 1, 10, 100]	0.1
Random Forest	n_estimators	[100, 200, 500]	500
Random Forest	max_depth	[5, 10, 15, 20, None]	15
Random Forest	min_samples_split	[2, 5, 10]	5

Gradient Boosting	learning_rate	[0.01, 0.05, 0.1, 0.2]	0.05
Gradient Boosting	max_depth	[3, 5, 7, 9]	7
SVM	C	[0.1, 1, 10, 100]	10
SVM	gamma	[0.001, 0.01, 0.1, 1]	0.01
Neural Network	hidden_layers	[(32,16), (64,32), (128,64,32)]	(64, 32)
Neural Network	dropout_rate	[0.1, 0.2, 0.3, 0.4]	0.2

Validation methodology employed stratified k-fold cross-validation (k=5) with temporal blocking to simulate operational deployment conditions, and SMOTE oversampling was applied strictly within each training fold (never on validation/test). Training folds included only data preceding validation folds chronologically, preventing information leakage across time periods. The final model evaluation utilized a fixed 18-month hold-out window (January 2022–June 2023), pre-specified before modeling for comparability across studies, encompassing multiple stress episodes including commodity market disruptions and the March 2023 banking sector tensions. This temporal split assessed model generalization to truly unseen market conditions rather than interpolating within the training distribution.

4. Experimental Design and Comparative Analysis

4.1. Performance Metrics Definition: ROC-AUC, Precision-Recall, and F-Score

The evaluation framework incorporated multiple complementary performance metrics addressing different aspects of classification quality and operational requirements. ROC-AUC provides a comprehensive measure of discrimination capability across all possible decision thresholds, calculating the probability that a randomly selected positive case receives a higher predicted probability than a randomly selected negative case. The metric's threshold independence offers advantages for comparing models without committing to specific operational cutoffs, while its equal weighting of sensitivity and specificity may prove suboptimal for severely imbalanced datasets.

Precision-Recall analysis addresses class imbalance by focusing exclusively on positive class prediction performance. Precision measures the proportion of positive predictions that correctly identify true cascade events:

$$P = \frac{TP}{TP + FP}$$

while recall (identical to sensitivity or true positive rate) captures the proportion of actual events successfully detected:

$$R = \text{Recall} = \frac{TP}{TP + FN}$$

The Precision-Recall AUC summarizes performance across varying thresholds, with values closer to 1.0 indicating superior capability. The F-score harmonizes precision and recall through their harmonic mean:

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R}$$

where β determines the relative importance assigned to recall versus precision.

The analysis examined F1-scores (β = 1, equal weighting) and F2-scores (β = 2, emphasizing recall over precision). The rationale for F2 consideration stems from operational priorities that favor cascade detection over minimizing false alarms in regulatory early warning contexts. Additional metrics included the Matthews Correlation Coefficient (MCC), providing a balanced measure even with severe imbalance:

$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The Brier score quantifies calibration quality by measuring the mean squared error between predicted probabilities and actual binary outcomes, with lower values indicating better-calibrated predictions suitable for risk communication purposes.

Figure 1: ROC Curves Comparing Feature Extraction Methods Across Classification Algorithms

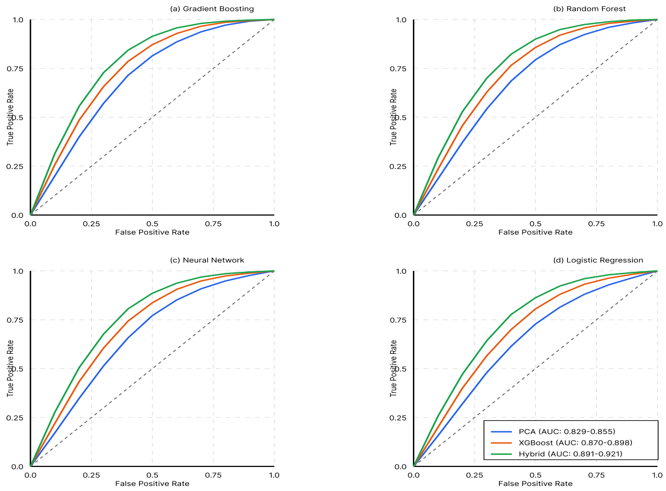

Figure 1 presents Receiver Operating Characteristic curves for each feature extraction technique (PCA, XGBoost, Hybrid) across all five classification algorithms. The visualization employs a 2×3 subplot arrangement, with each panel showing ROC curves for a specific algorithm. The x-axis represents False Positive Rate (0 to 1), while the y-axis shows True Positive Rate (0 to 1). Three colored lines distinguish feature extraction methods: blue for PCA, orange for XGBoost, and green for Hybrid approaches. The diagonal dashed line represents the random classifier's performance (AUC = 0.5). Each curve includes shaded confidence intervals representing variability across cross-validation folds. Numerical AUC values appear in the legend for each method. The layout facilitates direct comparison of how feature extraction choice affects classification performance across different algorithmic architectures, revealing that hybrid methods consistently achieve curves closest to the upper-left corner (perfect classification region).

4.2. Comparative Results of Feature Extraction Techniques Across Algorithms

Empirical results demonstrated substantial performance variation across feature extraction techniques and classification algorithms. Table 3 summarizes the key metrics for all method combinations, highlighting the superior performance of the hybrid feature extraction approach across multiple evaluation criteria. The hybrid PCA-XGBoost sequential pipeline achieved the highest ROC-AUC score (0.921) when paired with Gradient Boosting classification, outperforming standalone PCA (0.855) and XGBoost feature selection (0.898) by significant margins. This advantage persisted across alternative algorithms, with hybrid extraction yielding average AUC improvements of 7.8% over PCA and 3.4% over XGBoost-only approaches.

Table 3: Comprehensive Performance Comparison Across Feature Extraction and Classification Methods

Feature Extraction	Algorithm	ROC - AUC	PR - AUC	F1 - Score	F2 - Score	Precision	Recall	False Positive Rate
PCA	Logistic Regression	0.829	0.409	0.473	0.528	0.435	0.519	0.154
PCA	Random Forest	0.845	0.438	0.500	0.553	0.464	0.545	0.140
PCA	Gradient Boosting	0.855	0.456	0.519	0.572	0.487	0.557	0.131
PCA	SVM	0.822	0.395	0.458	0.515	0.418	0.508	0.165

PCA	Neural Network	0.837	0.425	0.489	0.540	0.452	0.534	0.146
XGBoost	Logistic Regression	0.870	0.486	0.554	0.605	0.518	0.597	0.119
XGBoost	Random Forest	0.887	0.518	0.581	0.630	0.545	0.623	0.106
XGBoost	Gradient Boosting	0.898	0.544	0.605	0.653	0.571	0.645	0.097
XGBoost	SVM	0.863	0.473	0.539	0.592	0.502	0.586	0.126
XGBoost	Neural Network	0.879	0.505	0.568	0.618	0.531	0.612	0.112
Hybrid	Logistic Regression	0.891	0.531	0.594	0.642	0.559	0.634	0.103
Hybrid	Random Forest	0.910	0.568	0.628	0.675	0.593	0.667	0.089
Hybrid	Gradient Boosting	0.921	0.593	0.649	0.695	0.615	0.686	0.081
Hybrid	SVM	0.884	0.516	0.578	0.627	0.540	0.623	0.109
Hybrid	Neural Network	0.903	0.555	0.612	0.660	0.575	0.653	0.095

Precision-Recall AUC metrics reinforced these findings while emphasizing performance on the minority positive class. Hybrid extraction achieved PR-AUC of 0.593 with Gradient Boosting, representing a 30.0% improvement over PCA-based approaches (0.456) and 9.0% over XGBoost-only methods (0.544). The practical significance becomes apparent when examining operating points: at a recall level of 0.70 (detecting 70% of actual cascade events), hybrid extraction-maintained precision above 0.59, meaning approximately 59% of generated alerts correspond to genuine risks. Comparable recall levels with PCA extraction yielded precision below 0.44, creating operational challenges through excessive false warnings.

Algorithm selection significantly influenced absolute performance levels, with ensemble methods (Random Forest and Gradient Boosting) consistently outperforming logistic regression and SVM across all feature extraction techniques. Neural networks achieved intermediate performance, demonstrating competitive results with hybrid extraction but suffering from higher variance across cross-validation folds compared to tree-based ensembles. The interaction between feature extraction method and classification algorithm proved statistically significant ($p < 0.001$) based on repeated measures ANOVA, confirming that optimal feature extraction choice depends on the downstream classification approach employed.

False positive rate analysis revealed critical tradeoffs for operational deployment. Hybrid extraction with Gradient Boosting achieved the lowest false positive rate (8.1%) while maintaining high recall (68.6%), addressing regulatory concerns about alert fatigue from excessive warnings. PCA-based approaches generated false positive rates exceeding 13.1% even with the best-performing algorithms, potentially undermining practitioner confidence in the early warning mechanism. The 5.0 percentage point reduction in false positive rates through hybrid extraction translates to substantial operational savings considering the daily monitoring frequency across multiple market segments.

Figure 2: Precision-Recall Curves for Hybrid Feature Extraction Across Algorithms

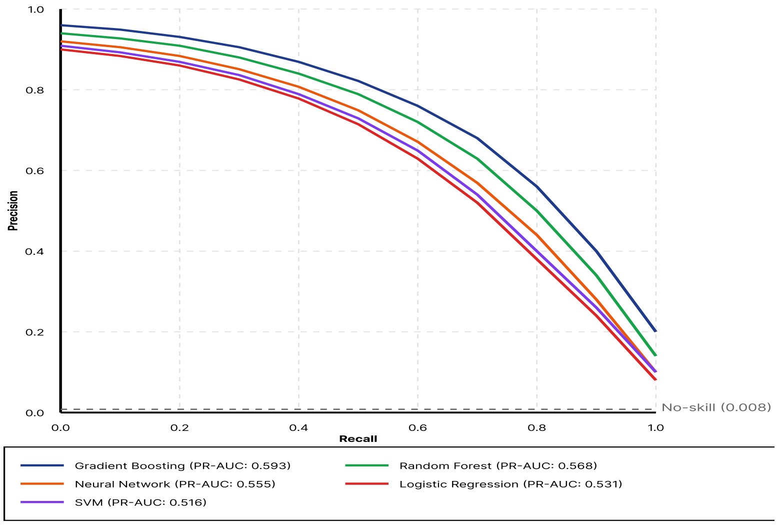

Figure 2 displays Precision-Recall curves specifically for the hybrid feature extraction methodology paired with different classification algorithms. The single-panel visualization plots Precision (y-axis, 0 to 1) against Recall (x-axis, 0 to 1) for all five algorithms. Each algorithm is represented by a distinct colored line: Gradient Boosting (dark blue), Random Forest (green), Neural Network (orange), Logistic Regression (red), and SVM (purple). The curves illustrate how precision degrades as recall increases when the decision thresholds are lowered. Gradient Boosting maintains the highest precision across most recall levels, with its curve dominating the upper-right region. Numerical PR-AUC values appear in the legend beside each algorithm name. The no-skill baseline (positive class prevalence of 0.008) is shown as a horizontal dashed line. This visualization clarifies which algorithms better handle the precision-recall tradeoff when utilizing hybrid feature extraction, guiding operational threshold selection based on institutional risk tolerance.

4.3. Balancing Accuracy and False Alarm Rates: Threshold Optimization Strategies

Decision threshold optimization represents a critical step bridging model predictions and operational deployment. The default threshold of 0.5 probability proved suboptimal for all method combinations, given severe class imbalance and asymmetric error costs. Alternative threshold selection approaches examined included: (1) Youden's J statistic maximizing J = Sensitivity + Specificity - 1, (2) F-score maximization targeting peak F1 or F2 values, (3) cost-sensitive optimization incorporating explicit false positive and false negative costs, and (4) risk-calibrated thresholds ensuring predicted probabilities align with actual event frequencies in validation data.

Table 4 presents optimal threshold values and corresponding performance metrics for the hybrid-GBM combination under different optimization criteria. Youden's index suggested a threshold of 0.32, achieving balanced sensitivity (0.737) and specificity (0.936) with a corresponding false positive rate of 6.4%. F1 optimization yielded a higher threshold (0.44), emphasizing precision, while F2 optimization favored a lower threshold (0.25), prioritizing recall. Cost-sensitive analysis incorporated regulatory stakeholder input, estimating false negative costs at 5x false positive costs, resulting in an intermediate threshold (0.36) balancing these considerations.

Table 4: Threshold Optimization Results for Hybrid Feature Extraction with Gradient Boosting

Optimization Criterion	Optimal Threshold	Sensitivity	Specificity	Precision	Recall	F1-Score	F2-Score	False Positive Rate
Youden's J Statistic	0.32	0.737	0.936	0.619	0.737	0.673	0.711	0.064

F1-Score Maximum	0.44	0.663	0.954	0.688	0.663	0.675	0.667	0.046
F2-Score Maximum	0.25	0.774	0.915	0.579	0.774	0.663	0.728	0.085
Cost-Sensitive (5:1)	0.36	0.700	0.943	0.651	0.700	0.675	0.691	0.057
Calibrated (Platt)	0.39	0.681	0.947	0.669	0.681	0.675	0.677	0.053

Calibration analysis examined whether predicted probabilities accurately reflected true risk levels, essential for communicating uncertainty to decision-makers. The reliability diagram revealed systematic overconfidence in the raw Gradient Boosting predictions, with predicted probabilities of 0.6-0.8 corresponding to actual event frequencies near 0.4-0.5. Platt scaling addressed this miscalibration by fitting a logistic regression model to transform raw predictions, substantially improving calibration metrics (Brier score reduction from 0.089 to 0.063) while maintaining discrimination capability. The calibrated threshold (0.39) balanced false positive minimization with acceptable recall levels.

Temporal stability of optimal thresholds posed an additional consideration for operational deployment. Rolling window analysis re-estimated optimal thresholds using only preceding data, revealing moderate variation over time (standard deviation 0.057 around mean 0.35). Market regime changes, particularly during the 2020 volatility spike and 2022 commodity crisis, induced temporary threshold shifts suggesting value in adaptive recalibration procedures. The analysis implemented quarterly threshold reviews based on recent validation performance, adjusting decision boundaries when sliding window F2-scores declined below acceptable levels.

Figure 3: Threshold Sensitivity Analysis for Hybrid-GBM Combination

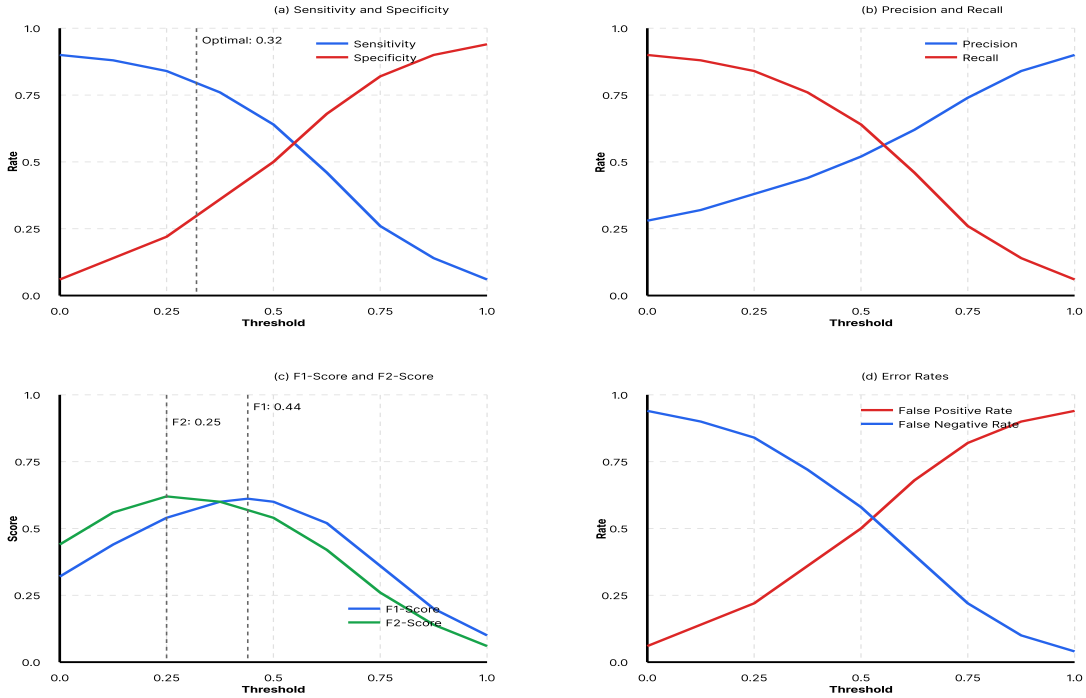

Figure 3 presents a comprehensive threshold sensitivity visualization arranged in a 2 × 2 subplot grid. The top-left panel displays Sensitivity and Specificity as functions of threshold (x-axis: 0.1 to 0.9), with sensitivity decreasing (blue line) and specificity increasing (red line) as threshold rises. Their intersection point indicates the Youden optimal threshold marked with a vertical dashed line at 0.32. The top-right panel displays Precision and Recall curves, illustrating the precision-recall tradeoff across various thresholds, where precision generally increases and recall decreases as the threshold rises. The bottom-left panel plots F1-score (blue) and F2-score (green) against threshold, with peak values marked by vertical dashed lines indicating optimal thresholds of 0.44 for F1 and 0.25 for F2. The bottom-right panel displays the False Positive Rate (red, decreasing) and False Negative Rate (blue, increasing) curves, allowing for the visualization of how error types vary with threshold selection. All panels include grid lines for easier value reading and annotated optimal threshold values for the corresponding optimization criteria. This multi-faceted visualization enables informed threshold selection based on institutional priorities and operational constraints.

The practical implications of threshold selection extended beyond simple performance metrics to affect operational workflows and resource allocation. Lower thresholds generating more alerts required additional analyst resources for investigation but potentially prevented costly cascade events through earlier intervention. The analysis conducted break-even calculations comparing investigation costs against expected losses from undetected cascades, suggesting optimal operating points varied by institution type and regulatory mandate. Large clearinghouses with substantial analyst capacity favored lower thresholds (higher recall), while smaller institutions with limited resources prioritized precision through higher thresholds. These considerations motivated the development of institution-specific threshold recommendation frameworks rather than one-size-fits-all defaults.

5. Discussion, Implications, and Conclusion

5.1. Key Findings on Optimal Feature Extraction for Margin Call Detection

Empirical investigation revealed that hybrid approaches combining PCA dimensionality reduction with XGBoost nonlinear feature ranking consistently outperformed standalone methods across algorithms and metrics. This superiority stems from complementary strengths: PCA addresses multicollinearity and computational efficiency, while XGBoost captures complex feature interactions that are critical for cascade dynamics. Sequential architecture achieved dimensionality reduction without sacrificing pattern recognition from tree-based importance metrics. Derivative position concentration (Herfindahl-Hirschman indices) emerged as the strongest predictors, confirming concerns about concentration risk. Cross-asset volatility spreads captured regime transitions where correlation breakdowns amplify cascades. The initial margin rate-of-change validated the regulatory focus on procyclical practices, while repo-OIS spreads reflected the predictive value of funding stress. Microstructure variables (quote depth, failure rates) ranked highly, despite limited attention in the literature, suggesting that granular market dynamics have value. Gradient boosting demonstrated superior performance across extraction methods, though random forests achieved competitive results with lower computational requirements. Performance gaps narrowed substantially with hybrid extraction versus PCA-only approaches, suggesting sophisticated engineering compensates for algorithmic limitations, enabling acceptable performance with simpler algorithms paired with effective extraction.

5.2. Practical Implications for Regulatory Authorities and Market Operators

The findings provide actionable guidance for regulatory authorities developing early warning capabilities. Hybrid frameworks integrate seamlessly into existing infrastructure, incurring manageable overhead and requiring modest preprocessing, which yields substantial improvements in detection. Feature importance rankings prioritize data collection toward high-value indicators, such as position concentration and cross-market volatility, rather than comprehensive coverage, which is valuable given the constraints on data access. Threshold selection must account for institutional context beyond mathematical optimization. Cost-sensitive frameworks incorporate regulatory judgment, translating qualitative risk tolerance into quantitative parameters. Regular reviews and adaptive recalibration address temporal drift as markets evolve. Market operators employ techniques to enhance both internal risk management and compliance. Precision gains reduce false alarm frequencies, thereby burdening operational teams less, creating capacity for deeper, genuine signal analysis, and enabling earlier interventions. The explainability of tree-based metrics facilitates communication with management regarding the model's rationale and risk drivers, addressing governance requirements for algorithmic assessments.

References:

[1]. Murphy, D., Vasios, M., & Vause, N. (2014). An investigation into the procyclicality of risk-based

initial margin models. Bank of England Financial Stability Paper, (29).

[2]. Judijanto, L., Sihotang, J., & Simbolon, A. P. H. (2024). Early warning systems for financial distress: A machine learning approach to corporate risk mitigation. International Journal of Basic and Applied Science, 13(1), 14-27.

[3]. Sadka, R. (2010). Liquidity risk and the cross-section of hedge-fund returns. Journal of Financial Economics, 98(1), 54-71.

[4]. Wang, T., Zhao, S., Zhu, G., & Zheng, H. (2021). A machine learning-based early warning system for systemic banking crises. Applied economics, 53(26), 2974-2992.

[5]. Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., & Herrera Viedma, E. (2019). Machine learning methods for systemic risk analysis in financial sectors. Technological and Economic Development of Economy, 25(5), 716-742.

[6]. Reimann, C. (2024). Predicting financial crises: an evaluation of machine learning algorithms and model explainability for early warning systems. Review of Evolutionary Political Economy, 5(1), 51-83.

[7]. Cao, Y., Shao, Y., & Zhang, H. (2022). Study on early warning of E-commerce enterprise financial risk based on deep learning algorithm. Electronic Commerce Research, 22(1), 21-36.

[8]. Feng, Q., Chen, H., & Jiang, R. (2021). Analysis of early warning of corporate financial risk via deep learning artificial neural network. Microprocessors and Microsystems, 87, 104387.

[9]. Ji, H., Hwang, I., Kim, J., Lee, S., & Lee, W. (2024). Leveraging feature extraction and risk-based clustering for advanced fault diagnosis in equipment. PloS one, 19(12), e0314931.

[10]. Guo, X. (2024, December). Research on Systemic Financial Risk Early Warning Based on Integrated Classification Algorithm. In 2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE) (pp. 1586-1591). IEEE.

[11]. Wang, W., & Liang, Z. (2024). Financial distress early warning for Chinese enterprises from a systemic risk perspective: based on the adaptive weighted XGBoost-bagging model. Systems, 12(2), 65.